

# A second-order stock market model

Robert Fernholz\*      Tomoyuki Ichiba<sup>†</sup>      Ioannis Karatzas<sup>‡</sup>

February 12, 2012

## Abstract

A *first-order model* for a stock market assigns to each stock a return parameter and a variance parameter that depend only on the *rank* of the stock. A *second-order model* assigns these parameters based on both the rank and the *name* of the stock. First- and second-order models exhibit stability properties that make them appropriate as a backdrop for the analysis of the idiosyncratic behavior of individual stocks. Methods for the estimation of the parameters of second-order models are developed in this paper.

*Key words:* stochastic portfolio theory, Atlas model, first-order model, second-order model.

*JEL Classification:* G10. *AMS 2010 Subject Classification:* 91B24.

## 1 Introduction

First-order and second-order stock market models are relatively simple stochastic models that manifest some of the stability properties of actual stock market behavior. These models are descriptive as opposed to normative, and are constructed using data analysis based on actual stock markets. First-order models are stock-market models where the parameters for return and volatility are based on the ranks of the stocks. These models were introduced in Fernholz (2002) and developed in Banner, Fernholz, and Karatzas (2005), and reflect the actual rank-based growth rates and variances of the stocks in the market. First-order models are asymptotically stable, and accurately reproduce the long-term characteristics of the market's capital distribution. However, these models are ergodic in the sense that each stock asymptotically spends equal average time at each rank, and this ergodicity property does not seem to be present in actual markets. This lack of verisimilitude is the motivation to consider the next level of complexity: second-order models.

Second-order models are a form of hybrid Atlas models, where the return and volatility parameters are based on the rank and the name (or index) of the stocks (see Ichiba et al. (2011)). While these models retain many of the characteristics of first-order models, the above ergodicity property is no longer present, and this produces a more realistic representation of actual stock market behavior. In second-order models, larger stocks tend to remain asymptotically among larger stocks, and

---

\*INTECH, One Palmer Square, Princeton, NJ 08542.

<sup>†</sup>Department of Statistics and Applied Probability, South Hall, University of California, Santa Barbara, CA 93106.

<sup>‡</sup>INTECH, One Palmer Square, Princeton, NJ 08542.

smaller stocks tend to remain among smaller stocks. This behavior is closer to that of actual stock markets, so second-order models provide a more accurate descriptive representation of stock market behavior.

Estimation of the parameters for first-order models is fairly straightforward, and can be accomplished without great ado. Second-order parameter estimation is somewhat more complicated. Here we shall focus on the growth-rate parameters, and find it necessary to rely on implicit methods to determine values for these parameters. Our purpose here is to develop techniques for estimating second-order growth-rate parameters, not to carry out an exhaustive examination of these parameters for an entire stock market. First, let us establish some formal definitions.

A *market* is a family of *stocks*  $X = (X_1, \dots, X_n)$  whose capitalizations are modeled by continuous, positive semimartingales that satisfy

$$d \log X_i(t) = G_i(t) dt + \sum_{\nu=1}^d S_{i\nu}(t) dB_\nu(t), \quad (1.1)$$

for  $t \in \mathbb{R}$ , where  $n \leq d$ ,  $B = (B_1, \dots, B_d)$  is an  $\mathbb{R}^d$ -valued Brownian motion defined on  $\mathbb{R}$ , and the  $G_i$  and  $S_{i\nu}$  are progressively measurable with respect to the Brownian filtration, with  $G_i$  locally integrable and  $S_{i\nu}$  locally square-integrable. The reason we define these processes on  $\mathbb{R}$  is that in practice we are confronted with time series over a given block of time, and the analysis of these series can be performed in both forward and reversed time. Hence, we see a sample in time of the processes  $X_1, \dots, X_n$  and draw our conclusions from this sample.

We shall assume that for any  $i \neq j$ , the intersection sets  $\{t : X_i(t) = X_j(t)\}$  have Lebesgue measure zero, almost surely, and we shall also assume that there are no *triple points*, i.e., if  $i < j < k$  then there is almost surely no  $t \in \mathbb{R}$  such that  $X_i(t) = X_j(t) = X_k(t)$ . The general setting for our model can be found in Fernholz (2002) and Fernholz and Karatzas (2009).

The value  $X_i(t)$  of the stock  $X_i$  at time  $t$  represents the total capitalization of the company at that time. If we let  $Z$  represent the total capitalization of the market, then

$$Z(t) \triangleq X_1(t) + \dots + X_n(t),$$

and we can define the *market portfolio* to be the portfolio  $\mu$  with weight processes given by the *market weights*

$$\mu_i(t) \triangleq \frac{X_i(t)}{Z(t)}, \quad \text{for } i = 1, \dots, n.$$

We shall assume that the market weight process  $\mu = (\mu_1, \dots, \mu_n)$  has a stable, or *steady-state*, distribution, and that the system is in that stable distribution. We shall be interested in the relative behavior of the log-capitalizations or log-weights. If  $\mu(t)$  is in its steady-state distribution, then the *log-difference processes* defined by

$$\log X_i(t) - \log X_j(t) = \log \mu_i(t) - \log \mu_j(t),$$

for  $i, j = 1, \dots, n$ , will also be in their steady-state distribution.

Consider the *ranked capitalization processes* corresponding to the  $X_i(t)$  in descending order

$$X_{(1)}(t) \geq \dots \geq X_{(n)}(t),$$

and the corresponding *ranked market weights*

$$\mu_{(1)}(t) \geq \dots \geq \mu_{(n)}(t).$$

Let  $r_t(i)$  represent the rank of  $X_i(t)$ , and let  $p_t$  be the inverse permutation of  $r_t$  (with ties in rank settled by the order of the indices), so

$$X_i(t) = X_{(r_t(i))}(t) \quad \text{and} \quad X_{(k)}(t) = X_{p_t(k)}(t),$$

and, similarly

$$\mu_i(t) = \mu_{(r_t(i))}(t) \quad \text{and} \quad \mu_{(k)}(t) = \mu_{p_t(k)}(t).$$

Hence,  $p_t(k)$  represents the index, or *name*, of the stock occupying rank  $k$  at time  $t$ .

The ranked market weights  $(\mu_{(1)}(t), \dots, \mu_{(n)}(t)) \equiv (\mu_{p_t(1)}(t), \dots, \mu_{p_t(n)}(t))$  comprise the *capital distribution curve* of the market at time  $t$ . The capital distribution curves over several decades of the 20th century can be seen in Figure 1, a version of which appears in Fernholz (2002). The curves in Figure 1 show the ranked market weights on December 31 of the years 1929, 1939, 1949, 1959, 1969, 1979, 1989, and 1999. During that period, the number of stocks in the market increased over each decade, so the decade associated with each curve is clear from the chart. We see that the capital distribution curve of the market shows a certain stability over time, so the assumption that  $\mu$  is in its steady state distribution would seem to be consistent with the observed data.

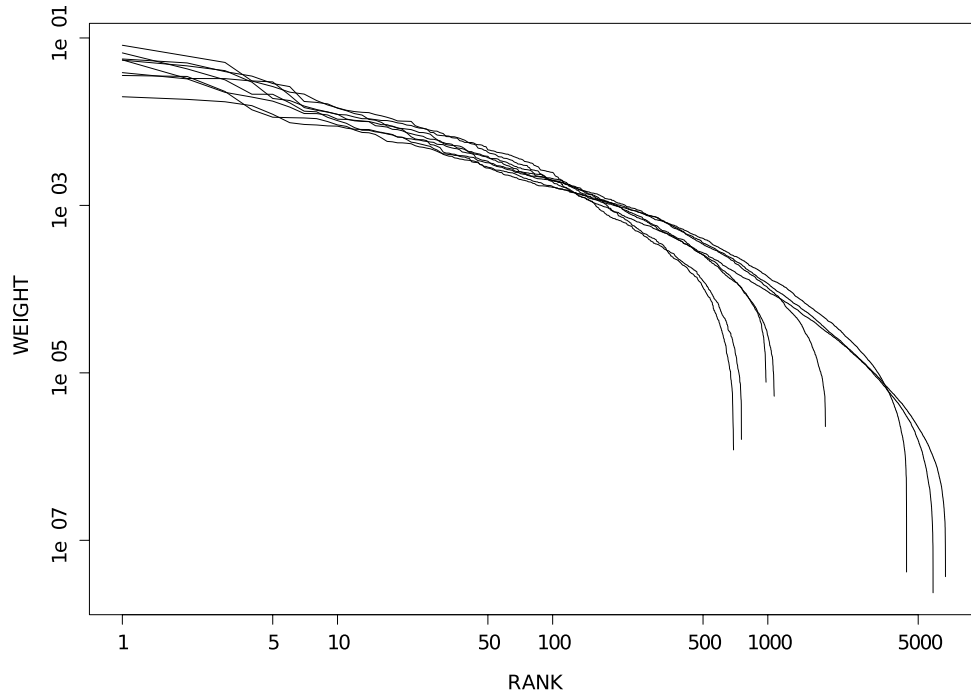


Figure 1: Capital distribution of the U.S. market: 1929–1999.  
The curves show the ranked weights at the end of each decade.

**Acknowledgements.** The authors are grateful to Adrian Banner, Daniel Fernholz, Vassilios Papathanakos, and Johannes Ruf for their many helpful discussions and suggestions, as well as for their participation and inspiration during the course of this research.

## 2 First-order models

A first-order model is a stock-market model in which each stock has constant growth and variance parameters that depend only on the *rank* of the stock by market capitalization. These models were developed in Fernholz (2002) and Banner et al. (2005), and can be constructed to reflect certain properties of actual stock markets. A first-order model that is based on an actual market will have a steady-state capital distribution curve that is about the same as the capital distribution curve for the actual market (see Fernholz (2002), Figure 5.6).

A *first-order model* is defined by a system  $\hat{X} = (\hat{X}_1, \dots, \hat{X}_n)$  of the form

$$\begin{aligned} d \log \hat{X}_i(t) &= g_{\hat{r}_t(i)} dt + \sigma_{\hat{r}_t(i)} dW_i(t), \\ &= \sum_{k=1}^n g_k \mathbf{1}_{\hat{r}_t(i)=k} = k dt + \sum_{k=1}^n \sigma_k \mathbf{1}_{\hat{r}_t(i)=k} = k dW_i(t), \end{aligned}$$

for  $i = 1, \dots, n$ , where  $g_1, \dots, g_n$  are real constants,  $\sigma_1, \dots, \sigma_n$  are positive constants, and  $(W_1, \dots, W_n)$  is an  $\mathbb{R}^n$ -valued Brownian motion, and where  $\hat{r}_t(i)$  represents the rank of  $\hat{X}_i(t)$  (analogously to  $r_t(i)$  for the rank of  $X_i(t)$ ). We shall assume that the  $g_k$  satisfy

$$g_1 + \dots + g_n = 0,$$

and

$$\sum_{k=1}^m g_k < 0,$$

for  $m < n$ . With these parameters, the  $\hat{X}_i$  form an *asymptotically stable* system, which means that the market weights  $\hat{\mu}_i(t) = \hat{X}_i(t) / (\hat{X}_1(t) + \dots + \hat{X}_n(t))$  satisfy

$$\lim_{t \rightarrow \infty} t^{-1} \log \hat{\mu}_i(t) = 0, \quad \text{for } i = 1, \dots, n,$$

and the limits corresponding to (2.1) and (2.3) below exist (see also Fernholz (2002), Definition 5.3.1).

Suppose we have a market  $X$ , and suppose that its market portfolio  $\mu$  is in the steady-state distribution. We define the asymptotic *rank-based relative variances* for the market by

$$\sigma_k^2 \triangleq \lim_{t \rightarrow \infty} t^{-1} \langle \log \mu_{(k)} \rangle(t), \quad (2.1)$$

and the asymptotic *rank-based relative growth rates* by

$$g_k \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \sum_{i=1}^n \mathbf{1}_{r_t(i)=k} = k d \log \mu_i(t), \quad (2.2)$$

and suppose these limits exist almost surely. Since these parameters are based on the market weight processes  $\mu_i$ , they represent values relative to the market portfolio  $\mu$ .

For  $k < \ell$ , let  $\Lambda k, \ell$  be the local time of the nonnegative semimartingale  $\log(\mu_{(k)}/\mu_{(\ell)}) \geq 0$  at the origin, and set  $\Lambda 0, 1 \equiv 0 \equiv \Lambda n, n+1$ . Since we have assumed that the  $X_i$  almost surely have no triple points, it follows that for  $\ell > k+1$ , the local time by  $\Lambda k, \ell$  is identically zero, so here we need to consider only local times of the form  $\Lambda k, k+1$ , and we have

$$d \log \mu_{(k)}(t) = \sum_{i=1}^n \mathbf{1}_{r_t(i)=k} = k d \log \mu_i(t) + \frac{1}{2} d \Lambda k, k+1(t) - \frac{1}{2} d \Lambda k-1, k(t), \text{ a.s.}$$

For  $k = 1, \dots, n-1$ , we can define the *asymptotic local time*

$$\lambda_{k,k+1} \triangleq \lim_{t \rightarrow \infty} t^{-1} \Lambda k, k+1(t), \quad (2.3)$$

which exists almost surely, and define  $\lambda_{0,1} \equiv 0 \equiv \lambda_{n,n+1}$ . It turns out that the estimation of the  $\lambda_{k,k+1}$  is not difficult, and the procedure is described in the appendix of Fernholz (2002). It can be shown (c.f. Proposition 5.3.2 in Fernholz (2002)) that

$$\mathbf{g}_k = \frac{1}{2}(\lambda_{k-1,k} - \lambda_{k,k+1}), \text{ a.s.} \quad (2.4)$$

holds for  $k = 1, \dots, n$ , and it follows that  $\mathbf{g}_1 + \dots + \mathbf{g}_n = 0$ .

The smoothed values of  $\sigma_k^2$  and  $\mathbf{g}_k$  for the largest 5120 stocks in the U.S. market for the decade 1990–1999 are shown in Figures 2 and 3. Since the number of stocks in the market changed during the decade of 1990–1999, we limit our attention here to the largest 5120 stocks, which is fewer than the number of stocks in the market at any time during that decade. The values in Figure 3 do not add up to zero, since the largest 5120 stocks are a strict subset of the larger market.

The first-order model  $\hat{X} = (\hat{X}_1, \dots, \hat{X}_n)$  such that

$$\begin{aligned} d \log \hat{X}_i(t) &= \mathbf{g}_{\hat{r}_t(i)} dt + \sigma_{\hat{r}_t(i)} dW_i(t), \\ &= \sum_{k=1}^n \mathbf{g}_k \mathbf{1}_{\hat{r}_t(i)} = k dt + \sum_{k=1}^n \sigma_k \mathbf{1}_{\hat{r}_t(i)} = k dW_i(t), \end{aligned}$$

where  $\hat{r}_t(i)$  is the rank of  $\hat{X}_i(t)$  at time  $t$ , is called the *first-order model* for the market  $X$ . As we have seen, the growth and variance parameters for  $\hat{X}$  are derived from the relative growth and variance parameters corresponding to the market weight processes  $\hat{\mu}_i$ , not directly from the capitalization processes  $\hat{X}_i$ .

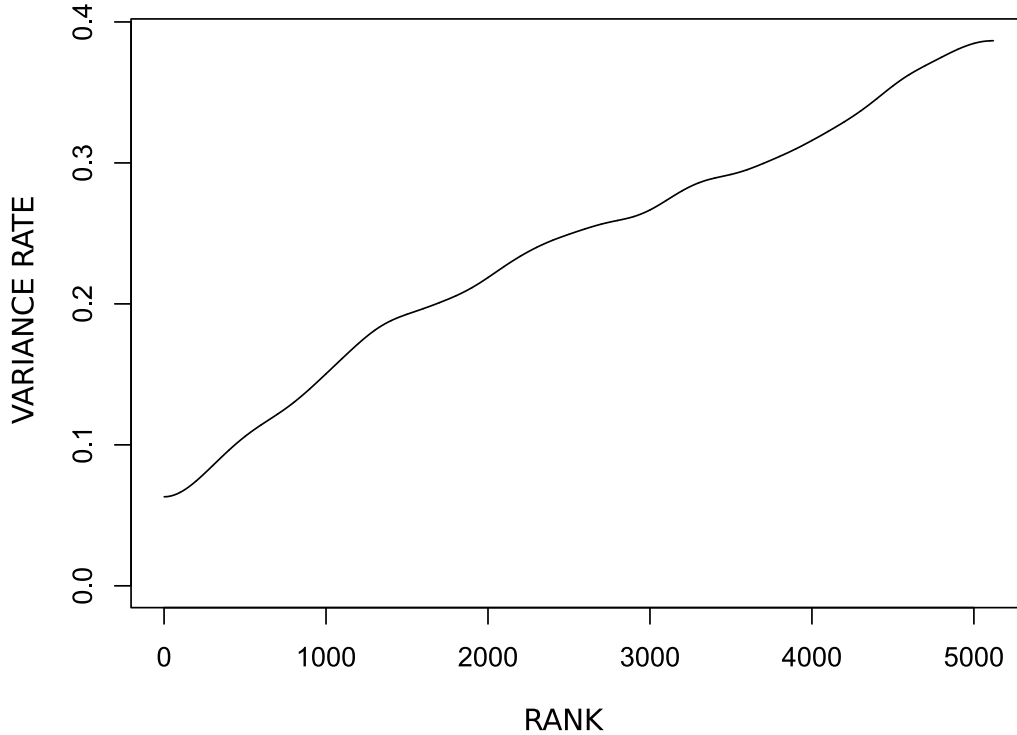


Figure 2: Smoothed values of  $\sigma_k^2$ ,  $k = 1, \dots, 5120$ , for U.S. market: 1990–1999.

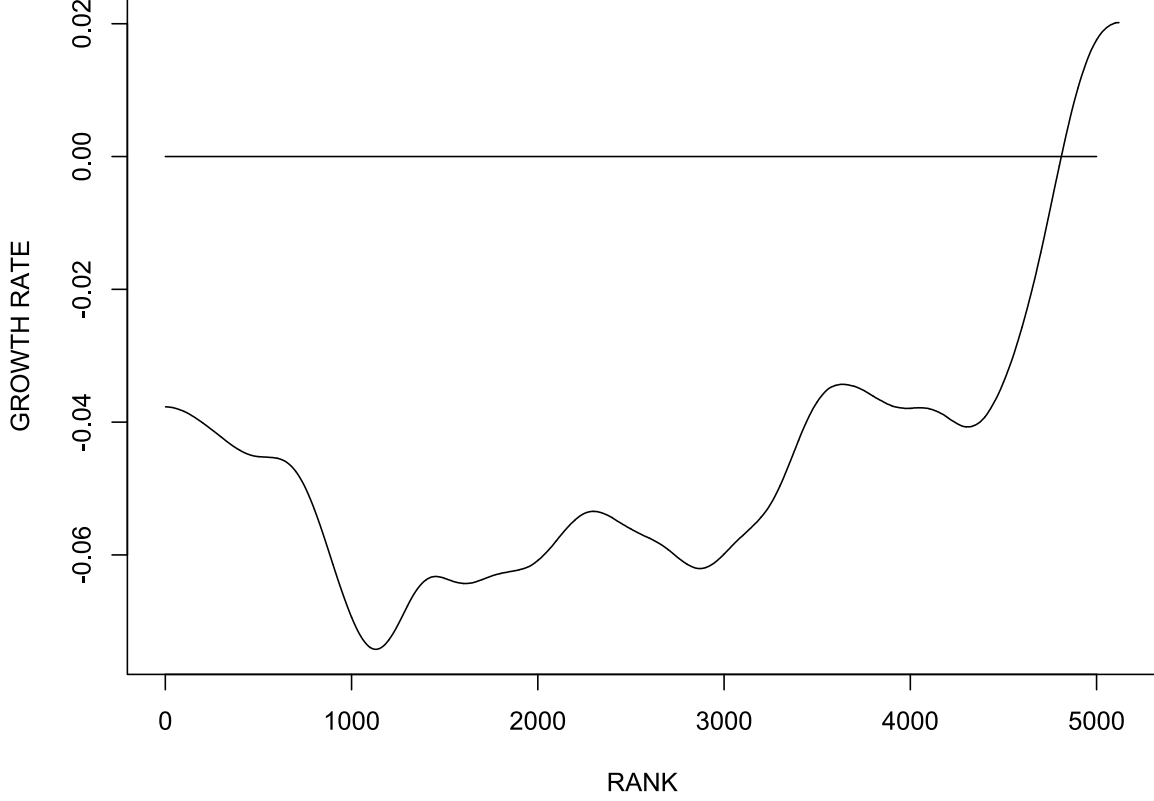


Figure 3: Smoothed values of  $\mathbf{g}_k$ ,  $k = 1, \dots, 5120$ , for U.S. market: 1990–1999.  
The  $\mathbf{g}_k$  satisfy  $\sum_{k=1}^n \mathbf{g}_k = 0$ , where  $n \cong 7000$ .

First-order models are ergodic in the sense that

$$\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \mathbb{1} \hat{r}_t(i) dt = k dt = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \mathbb{1} \hat{X}_i(t) dt = \hat{X}_{(k)}(t) dt = \frac{1}{n}, \text{ a.s.} \quad (2.5)$$

This ergodicity property does not seem to be present in real markets, but instead there exist *asymptotic occupation rates* defined by

$$\theta_{ki} \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \mathbb{1} r_t(i) dt = k dt = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \mathbb{1} X_i(t) dt = X_{(k)}(t) dt, \text{ a.s.} \quad (2.6)$$

Here  $\theta_{ki}$  represents the fraction of time that  $X_i$  spends in the  $k$ th rank. The  $n \times n$  matrix  $\theta = (\theta_{ki})$  is bistochastic, and we shall assume that all the entries are positive. For a first-order model, (2.5) implies that  $\theta_{ki} = 1/n$  for all  $i$  and  $k$ , and since this does not seem to characterize the behavior of real markets, we shall now consider a more general class of models.

### 3 Second-order models

A second-order model is a stock-market model in which each stock has constant growth and variance parameters that depend on the *rank* and *name*, or index, of the stock. Second-order models are

examples of *hybrid (Atlas) models*, which were discussed in Ichiba et al. (2011). A *second-order model* is defined by a system  $\widehat{X} = (\widehat{X}_1, \dots, \widehat{X}_n)$  of the form

$$\begin{aligned} d \log \widehat{X}_i(t) &= (\gamma_i + g_{\widehat{r}_t(i)})dt + \sigma_{i, \widehat{r}_t(i)} dW_i(t) \\ &= \left( \gamma_i + \sum_{k=1}^n g_k \mathbb{1}_{\widehat{r}_t(i)=k} \right) dt + \sum_{k=1}^n \sigma_{ik} \mathbb{1}_{\widehat{r}_t(i)=k} dW_i(t), \end{aligned} \quad (3.1)$$

for  $i = 1, \dots, n$ , with constants  $g_k$ ,  $\gamma_i$  and  $\sigma_{ik} > 0$ , for  $i, k = 1, \dots, n$ , and a Brownian motion  $W$ . In order for the  $\widehat{X}_i$  to be asymptotically stable, these parameters must satisfy

$$g_1 + \dots + g_n = 0 = \gamma_1 + \dots + \gamma_n,$$

and, for any permutation  $\pi \in \Sigma_n$ ,

$$\sum_{k=1}^m (g_k + \gamma_{\pi(k)}) < 0, \quad \text{for } m < n.$$

Here we are interested in estimating the growth-rate parameters  $\gamma_i$  and  $g_k$ . For simplicity, we shall consider only rank-based variances, and assume that  $\sigma_{ik}^2 = \sigma_k^2$  for all  $i$  and  $k$ .

It was shown in Ichiba et al. (2011) that a second-order model of the form (3.1) is asymptotically stable, and the asymptotic occupation rates

$$\widehat{\theta}_{ki} \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \mathbb{1}_{\widehat{r}_t(i)=k} dt \quad (3.2)$$

are defined for all  $i$  and  $k$ , almost surely. The matrix  $\widehat{\theta} = (\widehat{\theta}_{ki})$ , like  $\theta$  in (2.6), will be bistochastic with positive entries. We can generate the first-order parameters  $\widehat{\sigma}_k^2$  and  $\widehat{g}_k$  for  $\widehat{X}$  as in (2.1) and (2.2), with

$$\widehat{\sigma}_k^2 \triangleq \lim_{t \rightarrow \infty} t^{-1} \langle \log \widehat{\mu}_{(k)} \rangle(t),$$

and

$$\widehat{g}_k \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \sum_{i=1}^n \mathbb{1}_{\widehat{r}_t(i)=k} d \log \widehat{\mu}_i(t).$$

With these parameters, it was shown in Ichiba et al. (2011) that, almost surely,

$$\widehat{g}_k = g_k + \sum_{i=1}^n \widehat{\theta}_{ki} \gamma_i \quad (3.3)$$

$$0 = \gamma_i + \sum_{k=1}^n \widehat{\theta}_{ki} g_k. \quad (3.4)$$

In matrix form, this can be expressed

$$\begin{aligned} \widehat{g} &= g + \widehat{\theta} \gamma \\ 0 &= \gamma + \widehat{\theta}^T g, \end{aligned}$$

where  $\gamma$ ,  $g$ , and  $\widehat{g}$  are column vectors. From this we see that

$$\gamma = -\widehat{\theta}^T g, \quad (3.5)$$

so

$$\widehat{g} = (I_n - \widehat{\theta} \widehat{\theta}^T) g. \quad (3.6)$$

## 4 Estimation of second-order parameters

The first-order growth parameters  $\mathbf{g}_k$  for the market  $X$  can be estimated directly from the stock return time series; however, second-order growth parameters will have to be estimated indirectly. We wish to construct a second-order model that has first-order growth parameters equal to those of the market, and an occupation-rate matrix equal to the occupation-rate matrix  $\theta$  of the market. Under these circumstances, as in (3.6), we have

$$\mathbf{g} = (I_n - \theta\theta^T)\mathbf{g}, \quad (4.1)$$

and we wish to solve this equation for  $\mathbf{g}$ , the vector of name-based growth parameters for the second-order model of the market  $X$ . If we can solve (4.1) for this  $\mathbf{g}$ , then we can use (3.5), in the form

$$\gamma = -\theta^T \mathbf{g}, \quad (4.2)$$

to generate the name-based growth parameters  $\gamma_i$  for this second-order model. Let us first consider the matrix  $\theta$ .

The matrix  $\theta$  is bistochastic and we have assumed that all its entries are positive, so this also holds for  $\theta^T$  and  $\theta\theta^T$ . By the Perron-Frobenius theorem (see Perron (1907)), the symmetric matrix  $\theta\theta^T$  will have a simple eigenvalue equal to 1 with eigenvector  $e_1 = (1, 1, \dots, 1)'$ , and all the other eigenvalues will have absolute value less than 1. Hence,  $I_n - \theta\theta^T$  has rank  $n - 1$  and its kernel is generated by  $e_1$ , so the condition that the  $\mathbf{g}_k$  sum to zero means that  $\mathbf{g}$  is orthogonal to this kernel, and this ensures a unique solution to (4.1).

Unfortunately, it seems to be essentially impossible to estimate  $\theta$  with any reasonable accuracy, so although we can use this matrix to prove the existence and uniqueness of  $\mathbf{g}$ , in practice we cannot actually solve equation (4.1). Instead, let us consider (3.3) in the form

$$\mathbf{g}_k = \mathbf{g}_k + \sum_{i=1}^n \theta_{ki} \gamma_i. \quad (4.3)$$

We can use this equation to generate the  $\mathbf{g}_k$  recursively, and then estimate the  $\gamma_i$  from the returns data and the  $\mathbf{g}_k$ .

Let us assume that the market  $X$  is defined for all  $t \in \mathbb{R}$ , that the weight process  $\mu$  for  $X$  has a stable distribution, and that  $\mu$  is in that stable distribution. We can then define the *time-reversed* market  $\tilde{X}$  with stock capitalizations  $\tilde{X}_i(t) \triangleq X_i(-t)$  and weights  $\tilde{\mu}_i(t) \triangleq \mu_i(-t)$ , and with this definition we can define the expected backward occupation rates similarly to (2.6). Since the weight process is in its steady-state distribution, the limits of (2.6) will be the same at plus and minus infinity, so the forward and backward expected occupation rates  $\theta_{ki}$  will be equal. The results of Bertoin (1987) imply that the forward and backward asymptotic local times  $\Lambda_k, k + 1$  will also be the same, so the forward and backward versions of the  $\lambda_k$  are equal. Hence, it follows from (2.4) that the forward and backward  $\mathbf{g}_k$  are equal. In this case, (4.1) implies that the forward and backward values of the  $\mathbf{g}_k$  are equal, and from (4.2), we see that the forward and backward  $\gamma_i$  are also equal. Quadratic variation is invariant under time reversal, so the forward and backward  $\sigma_k$  will be the same. Hence, the first- and second-order models for  $X$  are the same as the corresponding models for  $\tilde{X}$ , and this allows us to use both  $X$  and  $\tilde{X}$  to estimate the second-order parameters.

In order to estimate the second-order parameters, it is necessary to observe the movement of market weights forward and backward in time. To this end, we define the concept of *flow* in a



market. The *forward flow*  $\phi_k$  of the market at rank  $k$  is defined for  $\tau \geq 0$  by

$$\phi_k(\tau) \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \log \left( \frac{\mu_{p_t(k)}(t + \tau)}{\mu_{(k)}(t)} \right) dt,$$

and the *backward flow*  $\tilde{\phi}_k$  of the market is defined by

$$\tilde{\phi}_k(\tau) \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \log \left( \frac{\tilde{\mu}_{p_t(k)}(t + \tau)}{\tilde{\mu}_{(k)}(t)} \right) dt.$$

In Figure 4 we see the exponential of forward and backward flow for the largest 250 stocks in the U.S. market over the decade from 1990 to 1999. The plots show the average exponential flow of each of the ten deciles of the top 250 stocks, with each decile comprising 25 stocks. The forward and backward flows need not be equal, and they do not appear to be equal in Figure 4. We see from the chart that for the largest 250 stocks the flow is downward. For the smaller stocks, we would expect the flow to be upward.

If we follow the flow of a stock that occupies a given rank at time zero, then the expected rank of the stock will change over time according to its flow. Suppose a stock is at rank  $k$  at time 0, and let us estimate its expected rank at time  $\tau \in \mathbb{R}$  by

$$\mathbf{R}_k(\tau) \triangleq \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T r_{s+\tau}(p_s(k)) ds.$$

In this case,  $\mathbf{R}_k(0) = k$ , and if this rank is among the higher ranks, we would expect the flow to be negative, which would mean that for  $\tau > 0$  we would expect that  $\mathbf{R}_k(\tau) \leq k$  and  $\mathbf{R}_k(-\tau) \leq k$ . We would like to use the  $\mathbf{R}_k$  to estimate the  $g_k$ , and although  $\mathbf{R}_k(\tau)$  need not equal  $\mathbf{R}_k(-\tau)$ , the  $g_k$  generated using either one will provide estimates for the solution of (3.6). Accordingly, we shall use the average of the two, with

$$\overline{\mathbf{R}}_k(\tau) \triangleq \left\lceil \frac{\mathbf{R}_k(\tau) + \mathbf{R}_k(-\tau)}{2} \right\rceil,$$

where the brackets signify the nearest integer. Values of  $\mathbf{R}_k(\tau)$  for  $k = 1, \dots, 250$  and  $\tau = \pm 4$  are shown in Figure 5, and the values for positive and negative  $\tau$  are clearly different. We have no explanation for this difference.

Figure 4 was generated by following the market weights of stocks that occupied a given rank at a given time in the decade from January 1, 1990 to December 31, 1999. Since stocks enter and leave the market, we used only the largest 250 stocks, after eliminating any stocks that did not have a full ten-year history. The trajectories of the weights for each of the top 250 ranks were followed for 1000 days forward or backward, and were then averaged over all starting dates that would allow the full 1000 days to be used. Finally, the ranks were separated into deciles, with ranks 1–25 in the first decile, 26–50 in the second decile, and so forth. The curves in Figure 4 represent the average trajectories for the weights of each of the ten deciles, forward and backward.

Figure 5 was generated by following the weight trajectories used for Figure 4 and, for each trajectory, noting the starting rank and ending rank, i.e., the rank after 1000 days (approximately four years of trading days). The final rank corresponding to the initial rank  $k$  in Figure 5 is the average ending rank for those trajectories that begin at rank  $k$  at time 0. This was carried out in forward time and reversed time.

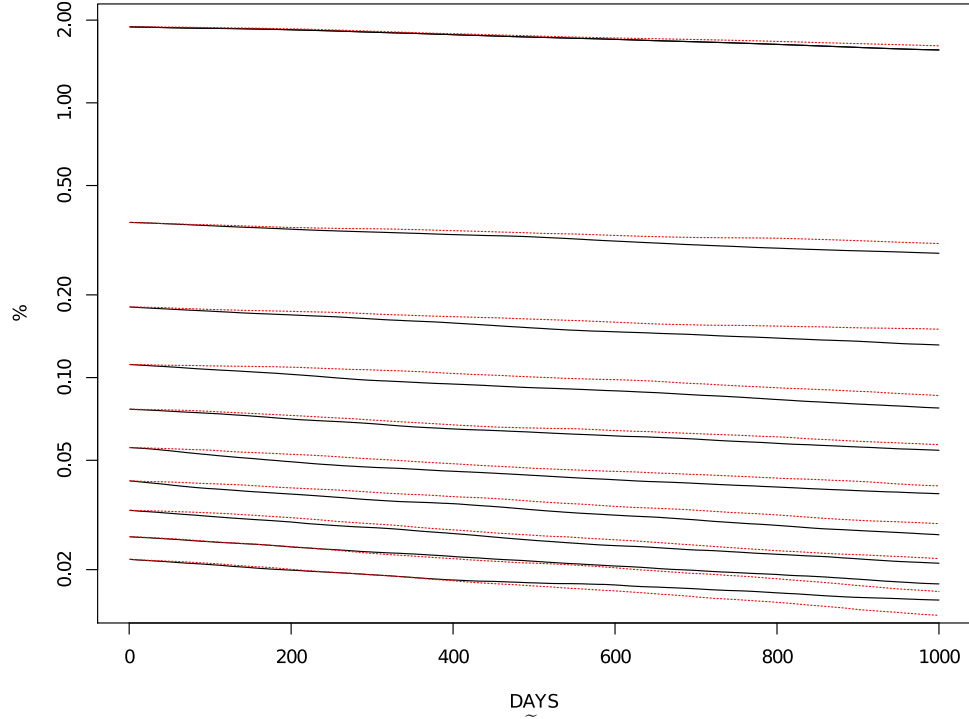


Figure 4:  $\mu_{(k)}(0)e^{\phi_k(\tau)}$  (black, solid),  $\mu_{(k)}(0)e^{\tilde{\phi}_k(\tau)}$  (red, dotted): 1990–1999. On average, a stock that starts at time 0 at rank  $k$  with market weight  $\mu_{(k)}(0)$  will move to weight  $\mu_{(k)}(0)e^{\phi_k(\tau)}$  at time  $\tau \in [0, 1000]$ , or to weight  $\mu_{(k)}(0)e^{\tilde{\phi}_k(\tau)}$  in reversed time.

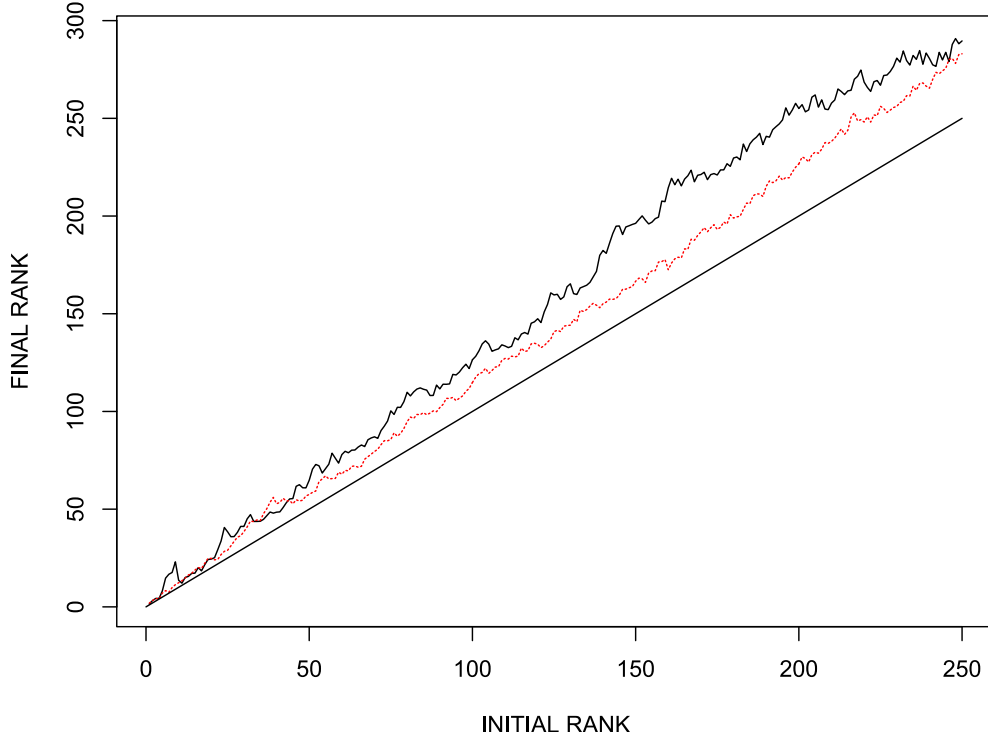


Figure 5:  $\mathbf{R}_k(4)$  (black, solid) and  $\mathbf{R}_k(-4)$  (red, dotted): 1990–1999. On average, a stock that starts a given initial rank will move to the corresponding final rank four years later, or earlier, in reversed time.

The straight line represents final rank = initial rank.

Let  $\mathbf{G}_k(\tau)$  be the expected growth rate at time  $\tau \in \mathbb{R}$  of a stock which occupies rank  $k$  at time 0, and we shall estimate  $\mathbf{G}_k(\tau)$  and  $\mathbf{G}_k(-\tau)$  from the slope of the forward and backward flow at rank  $k$ , so

$$\mathbf{G}_k(\tau) = D_\tau \phi_k(\tau) \quad \text{and} \quad \mathbf{G}_k(-\tau) = D_\tau \tilde{\phi}_k(\tau), \quad (4.4)$$

for  $\tau \geq 0$ , with  $\mathbf{G}_k(0) = g_k$ . We shall use the average

$$\overline{\mathbf{G}}_k(\tau) \triangleq \frac{1}{2}(\mathbf{G}_k(\tau) + \mathbf{G}_k(-\tau))$$

to estimate the rank-based growth rates  $g_k$ . In the data we analyzed, the derivatives in (4.8) at  $\tau = 4$  were estimated by measuring the rate of change of the flows  $\phi(\tau)$  and  $\tilde{\phi}(\tau)$  for the period from day 981 to day 1000, and then annualizing this rate.

Since for a given stock the name-based growth rate is invariant with rank, the same holds for the average of the name-based growth rates weighted by occupation rates,

$$\sum_{i=1}^n \hat{\theta}_{ki} \gamma_i.$$

Hence,

$$\overline{\mathbf{G}}_k(\tau) \cong g_{\overline{\mathbf{R}}_k(\tau)} + \sum_{i=1}^n \hat{\theta}_{ki} \gamma_i, \quad (4.5)$$

for  $\tau \in \mathbb{R}$ , where

$$g_{\overline{\mathbf{R}}_k(\tau)} \triangleq (\ell + 1 - \overline{\mathbf{R}}_k(\tau)) g_\ell + (\overline{\mathbf{R}}_k(\tau) - \ell) g_{\ell+1},$$

and  $\ell$  the largest integer such that  $\ell \leq \overline{\mathbf{R}}_k(\tau)$ . If we combine (4.5) with (3.3), we find that

$$g_{\overline{\mathbf{R}}_k(\tau)} \cong g_k + \overline{\mathbf{G}}_k(\tau) - g_k. \quad (4.6)$$

We can first estimate  $g_k$ ,  $\overline{\mathbf{G}}_k(\tau)$ , and  $\overline{\mathbf{R}}_k(\tau)$ , and then use (4.6) to recursively generate the values of the rank-based growth rates  $g_k$  for a subsequence of ranks of the form  $k, \overline{\mathbf{R}}_k(\tau), \overline{\mathbf{R}}_{\overline{\mathbf{R}}_k(\tau)}(\tau), \dots$ , as well as interpolated points.

Once we have estimates for the values of the  $g_k$ , we can estimate the  $\gamma_i$  directly by using

$$\gamma_i = \frac{1}{2} \left( \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T (d \log \mu_i(t) - g_{r_t(i)} dt) + \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T (d \log \tilde{\mu}_i(t) - g_{r_t(i)} dt) \right). \quad (4.7)$$

Our second-order model for the market  $X$  will then be

$$d \log \hat{X}_i(t) = (\gamma_i + g_{\hat{r}_t(i)}) dt + \sigma_{\hat{r}_t(i)} dW_i(t).$$

The various steps in the estimation process are shown in Figures 6, 7, and 8. In Figure 6 we see the estimated forward rank  $\overline{\mathbf{R}}_k(4)$  versus the initial rank  $k$ , and find that the relation is quite close to linear with

$$\overline{\mathbf{R}}_k(4) \cong 4.6 + 1.16k.$$

With this estimate, we can use (4.6) in the form

$$g_{(4.6+1.16k)} \cong g_k + \overline{\mathbf{G}}_k(4) - g_k \quad (4.8)$$

to estimate the  $g_k$  from the values of  $g_k$  and  $\overline{\mathbf{G}}_k(4)$ .

The values of  $\overline{\mathbf{G}}_k$  were estimated from the slopes of the flows used to generate Figure 4 for the ten rank-decile groups of 25 stocks each from the largest 250 stocks. Linear approximations for all the ranks were generated using a least squares fit. These results appear in Figure 7, and the corresponding linear equations are

$$\overline{\mathbf{G}}_k(0) = -4.2 - .034k \quad \text{and} \quad \overline{\mathbf{G}}_k(4) = -4.5 - .027k.$$

By using the values derived from these equations in (4.8) we can generate values for  $g_k$  for an increasing sequence of ranks  $k$ . The chart in Figure 8 shows these values with linear interpolation connecting the points to generate a continuous curve.

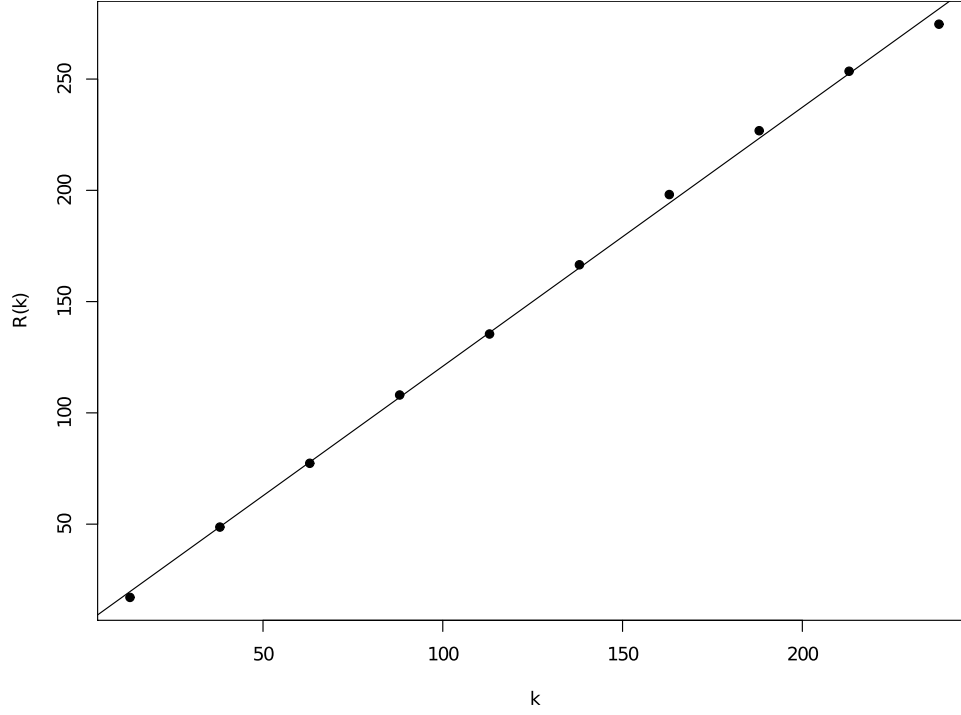


Figure 6: Estimated four-year forward rank  $\overline{\mathbf{R}}_k(4)$  corresponding to initial rank  $k$ .  
 $\overline{\mathbf{R}}_k(4) \cong 4.6 + 1.16k$ .

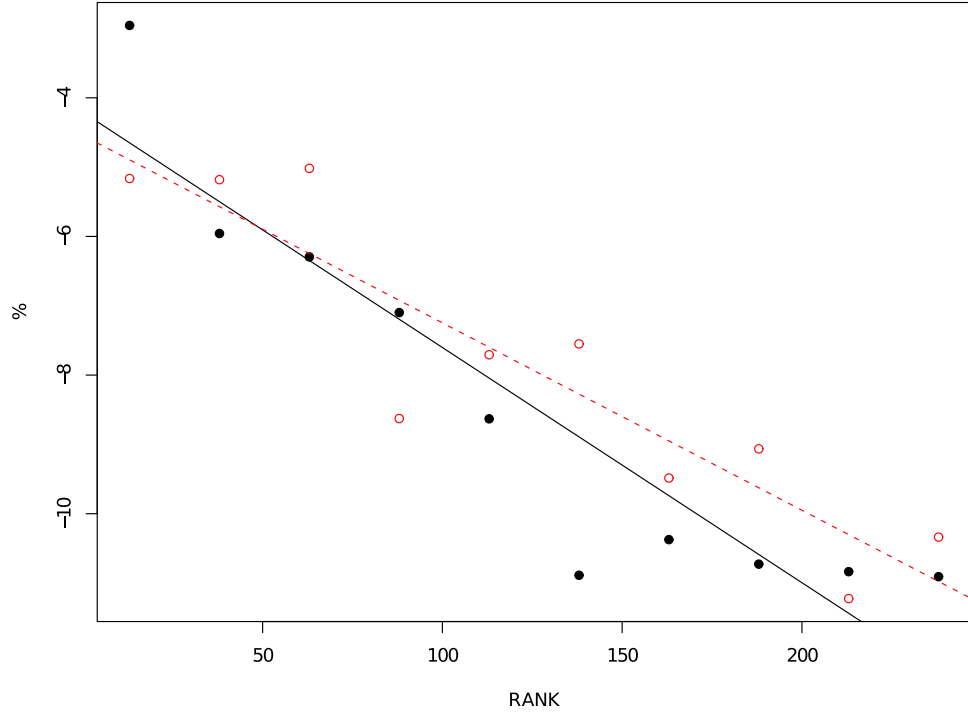


Figure 7: Estimated expected growth rates  $\overline{\mathbf{G}}_k(0)$  and  $\overline{\mathbf{G}}_k(4)$  at rank  $k$ .  
 $\overline{\mathbf{G}}_k(0) \cong -4.2 - .034k$  (black, solid line; dots),  
 $\overline{\mathbf{G}}_k(4) \cong -4.5 - .027k$  (red, broken line; circles).

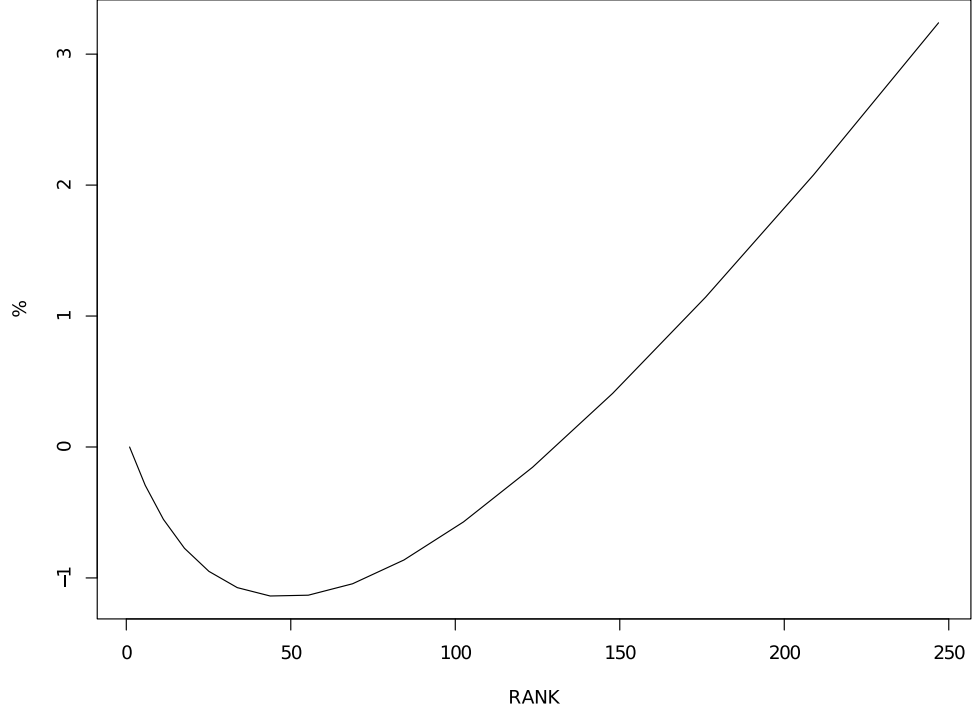


Figure 8: Values of  $g_k$  for ranks 1 to 250, calculated recursively and interpolated from  $g_{(4.6+1.16k)} \cong g_k + \overline{\mathbf{G}}_k(4) - \mathbf{g}_k$ . The  $g_k$  here are not normalized to add up to 0.

Once we have estimates for the values of the  $g_k$ , we can use these values along with (4.7) to estimate values for the  $\gamma_i$  of individual stocks by name. The integrals in (4.7) were approximated by daily logarithmic relative returns taken for each of the stocks along with the values of the  $g_k$ . The (non-normalized) values for  $\gamma_i$  for the decade 1990–1999 for a number of well-known stocks appear in Table 1. This is hardly a definitive study, so only a few stocks are included here. Moreover, in some future work, it would be desirable to have confidence intervals for these values, rather than point estimates. In that regard, probably the most promising method would use some form of jackknife estimator, with perhaps 12 pseudovalues generated by leaving out one month of the year at a time (see Mosteller and Tukey (1977)). Probably the entire estimation process would need to be repeated for each pseudovalue.

The values in Table 1 were estimated using combined forward and backward estimates, as in (4.7), for the decade 1990–1999. Using only forward estimates or only backward estimates for the  $\gamma_i$  could have produced biased estimates, since some of these companies grew considerably over that decade. For each company, the number in parentheses is the rank of the time-averaged log-weight of the stock during the decade.

While the values in Table 1 may not be definitive, they at least appear plausible. The higher-ranked stocks have generally higher  $\gamma$ , which should help them maintain their positions at the top of the market. At this writing, Apple, AAPL, has the highest market capitalization in the U.S. market, but in the 1990s we see that its average rank was 93, and its  $\gamma$  is correspondingly  $-1.67\%$ . Hence, the estimated  $\gamma_i$  provide no miraculous forecasts of future behavior; instead they reflect local stability consistent with the observed decade.

Table 1: Values of  $\gamma_i$  for various companies, 1990–1999.

Apple, AAPL (93)	$-1.67\%$
Coca Cola, KO (4)	$0.26\%$
Exxon, XON (3)	$0.11\%$
General Electric, GE (1)	$0.14\%$
International Business Machines, IBM (6)	$-0.10\%$
Microsoft, MSFT (5)	$-0.12\%$

## 5 Conclusion

The purpose of first- and second-order models for stock markets is to create a rigorous backdrop for the statistical analysis of the behavior of individual stocks. Second-order models provide a more accurate and complete representation of a stock market than is possible in first-order models. The estimation of parameters for second-order models is more involved than for first-order models, and implicit methods must be used. We have proposed methods for the estimation of second-order growth rate parameters, and with these methods a more complete stock-market model is possible. Nevertheless, our techniques are rudimentary, and we believe that future research will yield significant improvements.

## References

- Banner, A., R. Fernholz, and I. Karatzas (2005). On Atlas models of equity markets. *Annals of Applied Probability* 15, 2296–2330.
- Bertoin, J. (1987). Temps locaux et intégration stochastique pour les processus de Dirichlet. *Séminaire de Probabilités (Strasbourg)* 21, 191–205.
- Fernholz, R. (2002). *Stochastic Portfolio Theory*. New York: Springer-Verlag.
- Fernholz, R. and I. Karatzas (2009). Stochastic portfolio theory: an overview. In A. Bensoussan and Q. Zhang (Eds.), *Mathematical Modelling and Numerical Methods in Finance: Special Volume, Handbook of Numerical Analysis*, Volume XV, pp. 89–168. Amsterdam: North-Holland.
- Ichiba, T., V. Papathanakos, A. Banner, I. Karatzas, and R. Fernholz (2011). Hybrid Atlas models. *Annals of Applied Probability* 21, 609–644.
- Mosteller, F. and J. W. Tukey (1977). *Data Analysis and Regression*. Reading, MA: Addison Wesley.
- Perron, O. (1907). Zur theorie der matrices. *Math. Annalen* 64, 248–263.